DOCUMENT RESUME

ED 398 279 TM 025 482

AUTHOR Takahashi, Tomone; Nasser, Fadia
TITLE The Impact of Using Item Parcels on ad hoc Goodness of Fit Indices in Confirmatory Factor Analysis: An Empirical Example.
PUB DATE Apr 96
NOTE 29p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Arabic; Chi Square; Foreign Countries; *Goodness of Fit; Grade 10; High Schools; *High School Students; *Measurement Techniques; *Student Attitudes; *Test Anxiety
IDENTIFIERS Ad Hoc Groups; *Confirmatory Factor Analysis; Israeli Arabs; *Item Parcels

ABSTRACT

The Arabic version of I. G. Sarason's (1984) Reactions to Tests scale was used to examine the impact of using item parcels on ad hoc goodness of fit indices in confirmatory factor analysis. Item parcels with different numbers of items and different numbers of parcels per factor were used. Analyses were conducted on a sample of 420 tenth graders from 2 similar Israeli-Arab high schools. Model fit was examined in terms of chi square, chi square to degrees of freedom ratio, the goodness-of-fit index (GFI), the comparative fit index (CFI), the Tucker-Lewis index (TLI), and the single sample expected cross-validation index (ECVI). Model fit in terms of GFI, CFI, TLI, and ECVI systematically improved when the number of items per parcel increased. The P value associated with chi square indicated the same tendency. However, the chi square to degrees of freedom ratio did not change systematically. The analyses were repeated with another sample consisting of 374 tenth graders from the same 2 high schools. Basically, the results were the same as those of the first sample. (Contains 2 figures, 6 tables, and 20 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Running head: ITEM PARCELS

ED 398 279

The Impact of Using Item Parcels on ad hoc Goodness of Fit Indices in Confirmatory Factor Analysis: An Empirical Example

Tomone Takahashi[1]

Fadia Nasser

University of Georgia

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
TOMONE TAKAHASHI
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Paper presented at the annual meeting of the American Educational Research Association, April 8 - 12, 1996, New York.

---

[1] Correspondence should be directed to Tomone Takahashi, Department of Educational Psychology, 325 Aderhold Hall, University of Georgia, Athens, GA, 30602, E-mail: tomone@moe.coe.uga.edu

We want to thank Dr. Joseph Wisenbaker Dr. Jeri Benson, and Christine Distefano for their valuable comments on the earlier version of this paper.

BEST COPY AVAILABLE

Abstract

The Arabic version of Sarason's Reactions to Tests (RTT) scale was used to examine the impact of using item parcels on ad hoc goodness of fit indices in confirmatory factor analysis. Item parcels with different numbers of items and different numbers of parcels per factor were used. Analyses were conducted on a sample of 420 tenth graders from two similar Israeli-Arab high schools. Model fit was examined in terms of chi-square, chi-square to degrees of freedom ratio, the goodness-of-fit index (GFI), the comparative fit index (CFI), the Tucker-Lewis index (TLI), and the single sample expected cross-validation index (ECVI). Model fit in terms of GFI, CFI, TLI, and ECVI systematically improved when the number of items per parcel increased. The P value associated with chi-square indicated the same tendency. However, the chi-square to degrees of freedom ratio did not change systematically. The analyses were repeated with another sample consisting of 374 tenth graders from the same two high schools. Basically, the results were the same as those of the first sample.

Introduction

Item Parcels

Most questionnaires used in research on personality constructs consist of ordered categorical items. Usually, the method of maximum likelihood (ML) is used in factor analytic studies involving such items. This estimation method is based on the assumption that the data are continuous and normally distributed. These assumptions are frequently violated, especially when categorical items are analyzed, and the violation can result in misleading findings and conclusions about the factor structure (Bernstein & Teng, 1989; Kishton & Widaman, 1994).

One solution to this problem is to use item parcels. An item parcel is a simple sum of several items assessing the same construct (Kishton & Widaman, 1994; see Figure 1 for an example of parceling). This concept is also referred to as testlet (Wainer, & Kiely, 1987) or miniscale (Prats, 1990). By creating item parcels, new variables are constructed that are closer to continuous variables and allow for a distribution closer to normal. Thus, item parcels are more likely to meet the assumptions of maximum likelihood estimation than individual ordered-categorical items. In other words, parceling can be viewed as a heuristic approach to converting ordered-categorical data into more continuous data with an eye toward minimizing the size of the attenuation caused by using ordered-categorical variables. Thus, the primary reason for developing item parcels is to yield variance-covariance matrices that are amenable to a linear factor analysis (Schau, Stevens, Dauphinee, & Vecchio, 1995).

Formation of item parcels. Item parcels have been used in both personality and aptitude tests. Researchers have applied different methods to sum subsets of items to form item parcels. In the context of aptitude tests, item parcels have been constructed as one way of dealing with a variety of problems that might occur in the algorithmic construction of tests. Principal among these difficulties are problems with context effects, item-ordering, and content balancing (Wainer & Lewis, 1987). Item parcels have also been used for testing dimensionality of aptitude tests consisted of dichotomous items. For example, item parcels were created in a way such that they have approximately the same means and standard deviations. The parcels were formed from non-parallel items, which have different levels of difficulty and non-overlapping content (Cook, Dorans, & Eignor, 1988)

In the context of personality measures, item parcels have been used to improve psychometric properties of measures and to have a smaller number of observed variables. Item parcels have been constructed in several ways. Cattell and Burdsal (1975) primarily used the factor analysis method to form item parcels. To compare the performance of different methods of parceling in terms of model fit in confirmatory factor analysis (CFA), Kishton and Widaman (1994) compared unidimensional and domain representative parcels and recommended the latter method for improving the psychometric properties of behavioral measures of personality.

To our knowledge, only in one study, individual items were contrasted with item parcels in terms of their impact on overall model fit in CFA context. In that study, Prats (1990) used data

from the Test Anxiety Inventory to construct item parcels consisted of random combinations of two items. Her results were inconsistent in terms of model fit. That is, item parcels did not always perform better than individual items in terms of model fit.

Number of items per parcel. The effect of the number of items per parcel on the model fit has not been thoroughly investigated. In previous studies, the number of items per parcel in factor analytic methods were used varies from two (Prats, 1990) to eighteen (Kishton & Widaman, 1994). However, the impact of using different numbers of items per parcel has not been explored.

Evaluation of Model Fit

The impact of using parcels has been evaluated in terms of reliability, factor loadings, and model fit (Cook et al. 1988; Kishton & Widaman, 1994; Prats, 1990). Model fit in confirmatory factor analysis (CFA) is assessed by a variety of indices developed by several researchers (for a recent review of indices of model fit, see Hu & Bentler, 1995). Indices of fit are classified into two broad categories: absolute indices of fit which include chi-square, the Goodness of Fit Index (GFI), and the Expected Cross-Validation Index (ECVI); and three types of incremental indices of fit such as the Bentler and Bonnet's Normed Fit Index (NFI, Type I), the Tucker Lewis Index (TLI, Type II), and the Comparative Fit Index (CFI, Type III). There is no consensus among researchers with regard to the best combination of indices to evaluate model fit. However, Hoyle and Panter (1995) recommend using the absolute indices of fit in conjunction with a

representative of both the type II and III incremental indices to evaluate model fit.

The chi-square test. The conventional overall test of fit in covariance structure analysis assesses the magnitude of the discrepancy between the observed covariance matrix and the covariance matrix implied by the model. Under an assumed distribution and the hypothesized model $\Sigma(\theta)$ for the population covariance Σ, the test statistic $T = (N\text{-}1)F_{min}$ has an asymptotic (large sample) chi-square distribution. The statistic T is often denoted "the chi-square test."

In general, the H_0: $\Sigma = \Sigma(\theta)$ is rejected if the value of the T statistic exceeds a T_α based upon chi-square distribution at a given alpha level of significance. However, in the context of covariance structure analysis, the chi-square test is used as a descriptive device rather than a statistical test of fit. A chi-square value to its degrees of freedom ratio less than two suggests a reasonable model fit (Carmines, & McIver, 1981).

Because of the problems associated with chi-square test such as the sample size, statistical power, and violation of the multivariate normality assumption, the standard chi-square test may not be a sufficient guide to model fit. Also a chi-square test offers only a dichotomous decision strategy implied by a statistical decision rule and cannot be used to quantify the degree of fit along a continuum with some prespecified boundary. Thus, many so-called fit indices have been developed to assess the degree of congruence between the model and data (Bentler, 1990,

Bentler & Bonnet, 1980; Jöreskog & Sörbom, 1981 Tucker & Lewis, 1973).

Goodness of Fit Index (GFI). Jöreskog and Sörbom (1984) proposed GFI as a measure of the relative amount of variance and covariance in observed data that are accounted for by the covariance matrix of implied model. GFI carries an intuitive interpretation because it is analogous to the familiar R^2 value often reported in the context of multiple regression (Tanaka, 1993).

Tucker-Lewis Index (TLI). TLI is the classic index first developed by Tucker and Lewis (1973) under the assumption of normality and for the use of the ML estimation method uses information from the expected value of the chi-square of the target model. It was developed to quantify the degree to which a particular factor model is an improvement over a zero factor model when assessed by maximum likelihood. TLI is interpreted as a relative decrease in noncentrality per degrees of freedom. Thus, the TLI has an embodied parsimony component. Additionally, this index is not bounded by zero and one.

Comparative Fit Index (CFI). Similar to the TLI, the CFI (Bentler, 1990) can be interpreted as a comparative reduction in noncentrality with respect to a null model. According to Gerbing and Anderson (1993), CFI provides a more precise measure of fit than TLI because of its smaller empirical standard error. CFI is bounded by zero and one. Therefore, it is conceptually easier to interpret the CFI than to interpret the TLI.

Expected cross-validation Index (ECVI). The expected cross-validation index (ECVI) for a single sample as proposed by Browne and Cudeck (1989, 1993) reflects the expected discrepancy over all possible calibration samples. Smaller values of the ECVI indicate a higher probability that the model will be replicable across samples from the same population. For example, when the 90% confidence interval corresponding to ECVI includes zero, the model fits the data from future samples drawn from the same population 90% of the time.

TLI and CFI were chosen because they seem to be less influenced by sample size than other fit indices are (Hu & Bentler, 1995; Hoyle & Panter, 1995). Judging from the literature, the acceptable evaluation criteria for the listed indices are as follows: to indicate reasonable fit, chi-square to degrees of freedom-ratio should not exceed 2.00, and the values of GFI, TLI, and CFI should exceed .90. For ECVI smaller value indicates better model fit.

Purposes of the Study

Although item parcels have been used and some guidelines have been developed to aid researchers forming item parcels properly, to our knowledge, no studies in the literature investigated performance of different numbers of items per parcel. Therefore, the purposes of this study are to investigate (a) the impact of using item parcels vs. individual items and (b) the impact of the number of items per parcel on model fit as assessed by a variety of overall absolute and incremental fit indices.

Method

Subjects

In the present study, two samples were used. The first sample consists of 420 tenth graders aged 15 to 16 from two Arab high schools in the central district of Israel. Of the participants, 194 were male and 226 were female students. The second (replication) sample includes 374 (147 male and 227 female) students with similar characteristics to the students in the first sample.

Instrument

The Reaction to Test (RTT) scale developed by Sarason (1984) consists of 40 Likert-type items with a four-point response scale. The RTT scale used in this study was translated into Arabic by the second author. The theoretical structure of the RTT scale consists of four subscales referred to as 'worry,' 'tension,' 'test irrelevant thinking,' and 'bodily symptoms.' Each of the four original subscales includes ten items (Sarason, 1984). The four-factor model was used to compare models which consist of different numbers of parcels formed from different numbers of items.

Data Analysis

To compare the influence of different sizes and numbers of item parcels, four different item sets were analyzed. First, all 40 items were used to form two parcels consisted of five items for each factor and five parcels consisted of two items for each factor. Second, 36 items were selected based on content, item intercorrelations, and item reliability. These items were used to form three parcels per factor, each is consisted of three items.

Third, 32 items were selected based on the same criteria, and two and four parcels per factor were formed. Finally, 24 items were selected in the same manner to form two and three parcels per factor (see Figure 2).

Model fit was evaluated in terms of indices of fit obtained by the CALIS procedure in SAS 6.04. These indices include chi-square and its associated p value, chi-square to degrees of freedom ratio, GFI, TLI, CFI, and ECVI. All the analyses described above were repeated with a replication sample to test the stability of the results.

Results

First Sample

The results are summarized in terms of descriptive statistics, reliability (squared multiple correlation: SMC), and model fit across models using individual items vs. different sizes and different numbers of item parcels.

Descriptive statistics and the reliability of the individual items are shown in Table 1, and a sample of those corresponding to item parcels are shown in Table 2. Descriptive statistics of each model shows that the univariate normality of each indicator[2] in terms of skewness and kurtosis is improved when item parcels are used. As expected, the reliability (SMC) of item parcels is relatively higher than that of the individual items.

To examine the effect of the different numbers of indicators per subscale on its reliability (internal consistency), Cronbach's α was calculated for each of the four subscales of test anxiety

[2] Indicator refers to items or parcels.

(see Table 3). These result shows that the reliabilities did not drop when the number of indicators per subscale decreased (e.g., .80 in Model A Worry and .80 in Model C Worry).

Internal consistency (Cronbach's α) of each parcel was also examined in order to determine whether the item combinations used in this study were appropriate. When the number of items per parcel is greater than two (Models C, H, and K), Cronbach's alpha for each of the parcels is relatively high (.57 to .79), thus reflecting the internal consistency of the items in each parcel.

The results of CFA in the first sample are summarized in Table 3. When the same number of items was included in the model, model fit improved as a function of the number of items per parcel. GFI, CFI, TLI, and ECVI were improved when the number of items per parcel was increased. However, the improvement in the chi-square to degrees of freedom ratio is not systematic. When 40 and 36 items were analyzed, the smallest chi-square to degrees of freedom ratio corresponded to the models with individual items (Model A and Model D), while the models with the largest number of items per parcel yielded the largest chi-square to degrees of freedom ratios. However, when 32 and 24 items were analyzed, the smallest chi-square to degrees of freedom ratios were found in the models with the largest number of items per parcel (Model H and Model K), while the models with two items per parcel yielded the largest chi-square to degrees of freedom ratios (Model G and Model J).

Replication Sample

The tendency toward the improvement in normality and indicator reliability (SMC) when item parcels are used is similar to the one found in the first sample. However, the item level normality in the replication sample is somewhat better than its counterpart in the first sample (see Table 4 and 5).

The results of CFA for the replication sample are summarized in Table 6. Like the results in the first sample, GFI, CFI, TLI, and ECVI were improved when the number of items per parcel was increased. However, the changing pattern of chi-square to degrees of freedom ratio was not the same as the one found in the first sample. For example, in the first sample, the smallest chi-square to degrees of freedom ratio corresponds to model H (4 items per parcel, 2 parcels per factor) and model K (3 items per parcel, 2 parcels per factor), while the smallest chi-square to degrees of freedom ratio in the replication sample corresponds to model C (5 items per parcel and 2 parcels per factor).

Discussion and Conclusions

As expected, smaller skewness and kurtosis and higher indicator reliability (SMC) were found in item parcels than those in individual items. Also parcels with more items tend to have higher reliability than those with fewer items. Improvement in normality leads to improvement in reliability because one of the assumptions underlying reliability estimation is that scores of variables should be normally distributed.

The reliability (Cronbach's α) of the subscales did not drop when the number of the indicators per subscale decreased because

the information included in the subscale remains the same. Even though the number of indicators decreases, each of these indicators includes more variability as a result of extending the score range when more items are grouped in one parcel.

GFI, CFI, TLI, and ECVI improved systematically when the number of items per parcel increased. This improvement can be attributed to several factors. One factor is the decrease in the number of parameters to be estimated in each model. When the number of parameters to be estimated is small, the probability of making specification errors decreases. Hence, the standard error of estimation will be small. Improvements in continuity, normality, and indicator reliability may also be factors that contribute to the better fit of the model.

In sum, based on the limited design of the current study, it is safe to conclude that item parcels function better than individual items do when the maximum likelihood estimation method is used in CFA. The positive impact of using parcels on model fit is more obvious when some of the individual items depart from univariate normality and/or have low item reliability. When items are similar in content, parcels that consist of more items lead to better fit.

If researchers have a reason to believe that the set of items and the specified model represent the construct(s) of interest, it is recommended that they not delete items nor change model specification based on statistical criteria such as the modification index. An alternative way to improve model fit when ordered categorical items and maximum likelihood estimation are

used is to form parcels based on content similarity. It is more likely that the results from this method are replicable because this method is not data driven.

In the current study, the impact of using item parcels is limited to one personality measure. Therefore, further studies with other measures are needed to support the results of the present study. In addition, studies with simulated data are needed to confirm the results obtained from the empirical data used in this study.

Reference

Bentler, P. M. (1990). Comparative fit indices in structural models. Psychological Bulletin, 107, 238-246.

Bentler, P. M., & Bonnet, D. G. (1980). Significance tests and goodness-of-fit in the analysis of covariance structures. Psychological Bulletin, 88, 588-606

Bernstein, H., & Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. Psychological Bulletin, 105, 467-477.

Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. Multivariate Behavioral Research, 24, 445-455.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In Bollen, K. A., & Long, J. S. (Eds.), Testing structural equation models (pp. 136-162). Newbury Park, CA: Sage.

Carmines, E. G., & McIver, S. P. (1981). Analyzing models with unobserved variables: Analysis of covariance structures. In G. W. Bohrnstedt and E. F. Borgatta (Eds.), Social measurement: Current issues (pp. 65-115). Beverly Hills, CA: Sage.

Cattell, R. B., & Burdsal, C. A. (1975). The radial parcel double factoring design: A solution to item-vs-parcel controversy. Mutivariate Behavioral Research, 10, 165-179.

Cook, L. L., Dorans, N. J., & Eignor, D. R. (1988). An assessment of the dimensionality of three SAT-verbal test editions. Journal of Educational Statistics, 13, 19-43.

Gerbing, D. W., & Anderson, j. C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. In K. A. Bollen & J. S. Lond (Eds.), Testing structural equation models (pp. 40-65). Newbury Park, CA: Sage.

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.). Structural equation modeling: Concepts, issues, and applications (pp. 158-176). Thousand Oaks, CA: Sage.

Hu, L., & Bentler, P. M. (1995). Evaluation of model fit. In R. H. Hoyle (Ed.). Structural equation modeling: Concepts, issues, and applications (pp. 158-176). Thousand Oaks, CA: Sage.

Jöreskog, K. G., & Sörbom, D. (1984). LISREL V: Analysis of linear structural relationships by the method of maximum liklihood. Chicago: National Educational Resources.

Jöreskog, K. G., & Sörbom, D. (1984). LISREL VI user's guide (3rd ed.). Mooresville, IN: Scientific Software.

Kishton, J. M., & Widaman, K. F. (1994). Unidimensional versus domain representative parceling of questionnaire items: An empirical example. Educational and Psychological Measurement, 54, 757-765.

Prats, D. C. (1990, April). The effects of forming miniscales on the construct validity of the test anxiety inventory. Paper presented at The National Council on Measurement in Education, Boston.

Sarason, I. G. (1984). Stress, anxiety, and cognitive interference: Reactions to tests. Journal of Personality and Social Psychology, 46, 929-938.

Schau, C., Stevens, J., Dauphinee, T., & Veccio, A. D. (1995). The development and validation of the survey of attitudes toward statistics. Educational and Psychological Measurement, 55, 868-875.

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models (pp. 40-65). Newbury Park, CA: Sage.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. Psychometrika, 38, 1-10.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. Journal of Educational Measurement, 24, 185-201.

Table 1

Descriptives of Individual Items (Sample 1)

Item	Mean	SD	Skewness	Kurtosis	Item Reliabilities (SMC) 40	36	32	24
1	2.26	0.88	0.51	-0.28	.43	.43	.44	
2	2.19	1.07	0.47	-1.02	.21	.22	.22	.20
3	1.45	0.80	1.83	2.60	.30	.31	.30	.33
4	1.61	0.87	1.38	1.06	.29	.28	.27	.25
5	1.73	0.89	1.08	0.33	.18	.18		
6	2.13	1.08	0.53	-1.00	.34	.35	.35	
7	1.55	0.83	1.61	1.98	.34	.37	.38	.42
8	2.76	1.02	-0.17	-1.16	.30	.30	.31	.30
9	1.92	1.04	0.84	-0.53	.17	.17	.18	
10	1.26	0.68	2.69	6.48	.06			
11	2.62	1.06	0.03	-1.27	.40	.41	.41	
12	1.79	0.88	0.91	-0.01	.30	.31	.32	.35
13	2.54	1.08	-0.02	-1.28	.07			
14	1.45	0.81	1.83	2.57	.23	.22		
15	2.38	0.98	0.38	-0.85	.47	.48	.48	.47
16	2.53	1.03	0.14	-1.17	.43	.43	.43	.43
17	1.60	0.90	1.39	0.87	.21	.21	.23	.24
18	1.33	0.71	2.35	5.07	.43	.43	.39	.40
19	1.85	0.95	0.95	-0.05	.44	.45	.45	.49
20	2.58	1.02	0.05	-1.14	.38	.39	.41	.43
21	2.40	0.91	0.30	-0.71	.44	.44	.44	.42
22	2.14	1.08	0.52	-1.01	.27	.27	.27	.27
23	1.60	0.89	1.39	0.89	.53	.53	.55	.63
24	1.56	0.84	1.54	1.64	.50	.51	.53	.49
25	1.57	0.92	1.52	1.19	.43	.43	.43	.45
26	2.38	1.08	0.36	-1.12	.54	.53	.54	.52
27	2.35	1.00	0.25	-0.99	.47	.47	.47	.47
28	1.57	0.90	1.48	1.08	.27	.27	.26	
29	1.73	0.90	1.04	0.12	.36			
30	1.69	0.95	1.29	0.58	.27	.27	.27	.24
31	1.96	1.05	0.81	-0.58	.31	.31	.30	
32	1.35	0.75	2.23	4.13	.33	.35	.35	
33	2.14	1.04	0.54	-0.86	.48	.48	.48	.47
34	2.17	0.98	0.50	-0.72	.46	.46	.46	.46
35	2.35	0.95	0.39	-0.74	.44			
36	2.08	1.04	0.55	0.90	.34	.34		
37	2.16	1.06	0.54	-0.92	.41	.41	.40	.36
38	1.53	0.81	1.49	1.50	.37	.36		
39	1.93	0.96	0.75	-0.43	.19	.19	.19	
40	2.15	1.07	0.15	-1.23	.48	.47	.48	.48

Table 2

Descriptive Statistics of Item Parcels (Sample 1)

Model C (5 items per parcel, 2 parcels per subscale)

Parcel	Mean	SD	Skew-ness	Kur-tosis	Parcel Reliability (SMC)	Internal Consistency (Cronbach α)
W1	11.22	3.57	0.42	-0.54	.81	.75
W2	11.00	3.15	0.09	-0.56	.56	.58
TEN1	12.78	3.78	0.38	-0.73	.83	.79
TEN2	10.77	3.41	0.47	-0.32	.75	.72
TIT1	7.59	2.88	1.57	2.46	.63	.75
TIT2	7.86	2.88	1.31	1.51	.74	.70
BS1	8.42	3.31	1.16	0.69	.66	.76
BS2	9.34	3.02	0.71	0.02	.73	.65

Model H (4 items per parcel, 2 parcels per subscale)

Parcel	Mean	SD	Skew-ness	Kur-tosis	Parcel Reliability (SMC)	Internal Consistency (Cronbach α)
W1	8.84	2.87	0.47	-0.46	.67	.71
W2	9.45	2.76	0.14	-0.59	.61	.61
TEN1	9.43	3.17	0.36	-0.80	.78	.76
TEN2	9.04	3.01	0.45	-0.44	.75	.72
TIT1	5.90	2.31	1.71	3.23	.58	.67
TIT2	6.29	2.39	1.20	1.00	.74	.66
BS1	6.97	2.92	1.15	0.65	.62	.76
BS2	8.08	2.81	0.66	-0.09	.72	.74

Model K (3 items per parcel, 2 parcels per subscale)

Parcel	Mean	SD	Skew-ness	Kur-tosis	Parcel Reliability (SMC)	Internal Consistency (Cronbach α)
W1	6.92	2.35	0.37	-0.59	.65	.74
W2	7.53	2.33	0.04	-0.75	.51	.61
TEN1	7.30	2.50	0.31	-0.86	.71	.74
TEN2	6.78	2.44	0.36	-0.21	.69	.65
TIT1	4.33	1.84	1.74	2.95	.48	.68
TIT2	4.94	1.98	1.08	0.69	.75	.62
BS1	5.01	2.31	1.26	0.87	.56	.78
BS2	5.46	2.12	0.91	0.41	.63	.57

Table 3

Model Comparison (Sample 1)

	Number of items per indicator	Number of indicators per factor	χ^2	df	p	χ^2/df	GFI	CFI	TLI	ECVI (90% CI)	Subscale Reliability			
											W	Ten	TIT	BS
A.	1	10	1544.42	734	.001	2.10	.84	.86	.85	4.13 (3.86, 4.42)	.80	.87	.83	.83
B.	2	5	353.16	164	.001	2.15	.92	.95	.94	1.07 (0.95, 1.21)	.77	.86	.81	.81
C.	5	2	33.10	14	.003	2.36	.98	.99	.98	0.19 (0.15, 0.24)	.80	.88	.81	.82
D.	1	9	1265.65	588	.001	2.15	.86	.87	.86	3.42 (3.18, 3.68)	.81	.85	.83	.83
E	3	3	118.17	48	.001	2.46	.96	.97	.96	0.43 (0.36, 0.52)	.78	.85	.79	.82
F.	1	8	1035.67	458	.001	2.26	.87	.88	.87	2.83 (2.61, 3.07)	.79	.85	.80	.83
G.	2	4	236.88	98	.001	2.42	.93	.95	.94	0.75 (0.65, 0.87)	.75	.85	.77	.82
H.	4	2	19.28	14	.155	1.38	.99	.99	.99	0.15 (0.00, 0.19)	.78	.87	.79	.80
I.	1	6	479.42	246	.001	1.95	.91	.93	.92	1.41 (1.27, 1.58)	.78	.82	.77	.79
J.	2	3	120.14	48	.001	2.50	.95	.97	.95	0.43 (0.37, 0.52)	.75	.82	.75	.77
K.	3	2	19.36	14	.152	1.38	.99	.99	.99	0.15 (0.00, 0.19)	.73	.82	.75	.74

Note. GFI = Goodness of fit index; CFI = Comparative fit index;

TLI = Tucker-Lewis index; ECVI = Expected cross-validation index;

CI = Confidence interval; W = Worry; Ten = Tension; TIT = Test irrelevant thinking;

BS = Bodily symptoms.

Table 4

Descriptives of Individual Items (Sample 2)

Item	Mean	SD	Skewness	Kurtosis	Item Reliabilities (SMC) 40	36	32	24
1	2.51	0.80	0.55	-0.47	.43	.44	.44	
2	2.31	0.99	0.22	-0.99	.28	.28	.28	.26
3	1.66	0.90	1.23	0.52	.39	.41	.43	.45
4	1.78	0.89	0.97	0.11	.30	.30	.29	.29
5	1.99	0.97	0.71	-0.48	.33	.33		
6	2.20	1.00	0.40	-0.89	.36	.36	.36	
7	1.67	0.86	1.17	0.54	.43	.45	.47	.49
8	2.96	0.93	-0.38	-0.91	.37	.37	.39	.30
9	2.07	1.01	0.51	-0.89	.27	.29	.30	
10	1.33	0.69	2.40	5.65	.15			
11	2.90	0.98	-0.33	-1.06	.46	.46	.45	
12	1.89	0.91	0.81	-0.19	.42	.42	.44	.47
13	2.72	1.03	-0.25	-1.11	.04			
14	1.80	1.04	1.00	-0.35	.26	.25		
15	2.57	0.94	0.15	-0.95	.57	.58	.58	.58
16	2.75	0.93	-0.08	-0.99	.49	.51	.52	.53
17	1.51	0.86	1.56	1.34	.21	.21	.20	.21
18	1.41	0.76	1.93	3.11	.46	.45	.41	.38
19	1.92	0.97	0.87	-0.22	.45	.45	.45	.50
20	2.75	0.98	-0.14	-1.10	.46	.45	.48	.52
21	2.40	0.85	0.21	-0.55	.33	.33	.33	.33
22	2.54	1.12	-0.00	-1.36	.21	.21	.21	.22
23	1.64	0.88	1.27	0.72	.41	.41	.41	.48
24	1.62	0.88	1.31	0.81	.51	.54	.53	.52
25	1.53	0.85	1.56	0.57	.29	.29	.29	.33
26	2.74	1.01	-0.03	-1.28	.52	.51	.51	.52
27	2.93	0.93	-0.31	-0.99	.41	.41	.42	.43
28	1.89	1.05	0.87	-0.53	.27	.25	.26	
29	1.92	0.98	0.70	-0.64	.41			
30	1.71	0.94	1.13	0.17	.26	.26	.26	.24
31	1.95	0.99	0.73	-0.58	.42	.42	.42	
32	1.46	0.82	1.82	2.41	.35	.34	.33	
33	2.20	1.03	0.45	-0.91	.45	.44	.44	.45
34	2.36	0.95	0.35	-0.79	.41	.41	.41	.41
35	2.05	0.99	0.55	-0.79	.34			
36	1.92	1.03	0.82	-0.56	.37	.37		
37	2.15	1.01	0.49	-0.85	.37	.37	.37	.32
38	1.58	0.84	1.35	1.01	.48	.48		
39	2.10	0.98	0.52	-0.72	.27	.27	.26	
40	2.90	1.02	-0.37	-1.12	.48	.49	.49	.48

Table 5

Descriptive Statistics of Item Parcels (Sample 2)

Model C (5 items per parcel, 2 parcels per subscale)

Parcel	Mean	SD	Skew-ness	Kur-tosis	Parcel Reliability (SMC)	Internal Consistency (Cronbach α)
W1	12.50	3.34	0.23	-0.80	.80	.74
W2	12.84	3.09	0.04	-0.75	.64	.62
TEN1	12.31	3.64	0.18	-0.80	.78	.80
TEN2	12.14	3.45	0.31	-0.63	.77	.73
TIT1	8.25	3.21	1.13	0.75	.66	.79
TIT2	8.37	3.07	1.08	0.79	.89	.71
BS1	8.84	3.39	0.99	0.37	.61	.76
BS2	9.86	3.03	0.60	-0.06	.74	.69

Model H (4 items per parcel, 2 parcels per subscale)

Parcel	Mean	SD	Skew-ness	Kur-tosis	Parcel Reliability (SMC)	Internal Consistency (Cronbach α)
W1	9.75	2.70	0.25	-0.65	.72	.69
W2	10.13	2.73	0.12	-0.70	.73	.66
TEN1	10.26	3.05	0.15	-0.84	.74	.79
TEN2	10.16	2.88	0.18	-0.76	.77	.69
TIT1	6.63	2.61	1.21	1.37	.67	.70
TIT2	6.48	2.48	1.12	0.89	.78	.68
BS1	7.05	2.83	1.01	0.38	.56	.76
BS2	8.53	2.76	0.47	-0.45	.67	.76

Model K (3 items per parcel, 2 parcels per subscale)

Parcel	Mean	SD	Skew-ness	Kur-tosis	Parcel Reliability (SMC)	Internal Consistency (Cronbach α)
W1	7.68	2.12	0.14	-0.74	.65	.66
W2	8.02	2.26	-0.03	-0.83	.60	.67
TEN1	8.06	2.42	0.05	-0.94	.67	.79
TEN2	7.65	2.40	0.12	-0.90	.68	.63
TIT1	4.74	2.02	1.23	1.30	.54	.71
TIT2	5.02	2.00	0.94	0.36	.88	.62
BS1	5.10	2.19	1.05	0.39	.49	.74
BS2	5.63	2.11	0.79	0.17	.58	.59

Table 6

Model Comparison (Sample 2)

	Number of items per indicator	Number of indicators per factor	χ^2	df	p	χ^2/df	GFI	CFI	TLI	ECVI (90% CI)	Subscale Reliability			
											W	Ten	TIT	BS
A.	1	10	1397.27	734	.001	1.90	.84	.88	.87	4.26 (3.98, 4.57)	.82	.87	.86	.83
B.	2	5	345.57	164	.001	2.11	.92	.95	.94	1.88 (1.05, 1.35)	.79	.87	.84	.82
C.	5	2	22.26	14	.073	1.59	.96	.99	.99	0.18 (0.00, 0.23)	.83	.87	.87	.80
D.	1	9	1174.42	588	.001	2.00	.85	.88	.88	3.61 (3.35, 3.90)	.83	.86	.85	.83
E	3	3	138.83	48	.001	2.89	.94	.96	.95	0.54 (0.45, 0.65)	.82	.85	.84	.81
F.	1	8	982.26	458	.001	2.14	.85	.88	.87	3.05 (2.81, 3.31)	.81	.86	.82	.82
G.	2	4	232.66	98	.001	2.37	.93	.95	.94	0.84 (0.73, 0.97)	.78	.86	.80	.80
H.	4	2	30.41	14	.007	2.17	.98	.99	.99	0.20 (0.17, 0.27)	.84	.86	.84	.76
I.	1	6	532.38	246	.001	2.16	.90	.91	.90	1.74 (1.57, 1.93)	.79	.82	.81	.76
J.	2	3	113.11	48	.001	2.36	.95	.97	.95	0.47 (0.40, 0.57)	.77	.83	.80	.71
K.	3	2	33.51	14	.002	2.39	.98	.99	.97	0.21 (0.18, 0.27)	.77	.81	.82	.69

Note. GFI = Goodness of fit index; CFI = Comparative fit index;

TLI = Tucker-Lewis index; ECVI = Expected cross-validation index;

CI = Confidence interval; W = Worry; Ten = Tension; TIT = Test irrelevant thinking;

BS = Bodily symptoms.

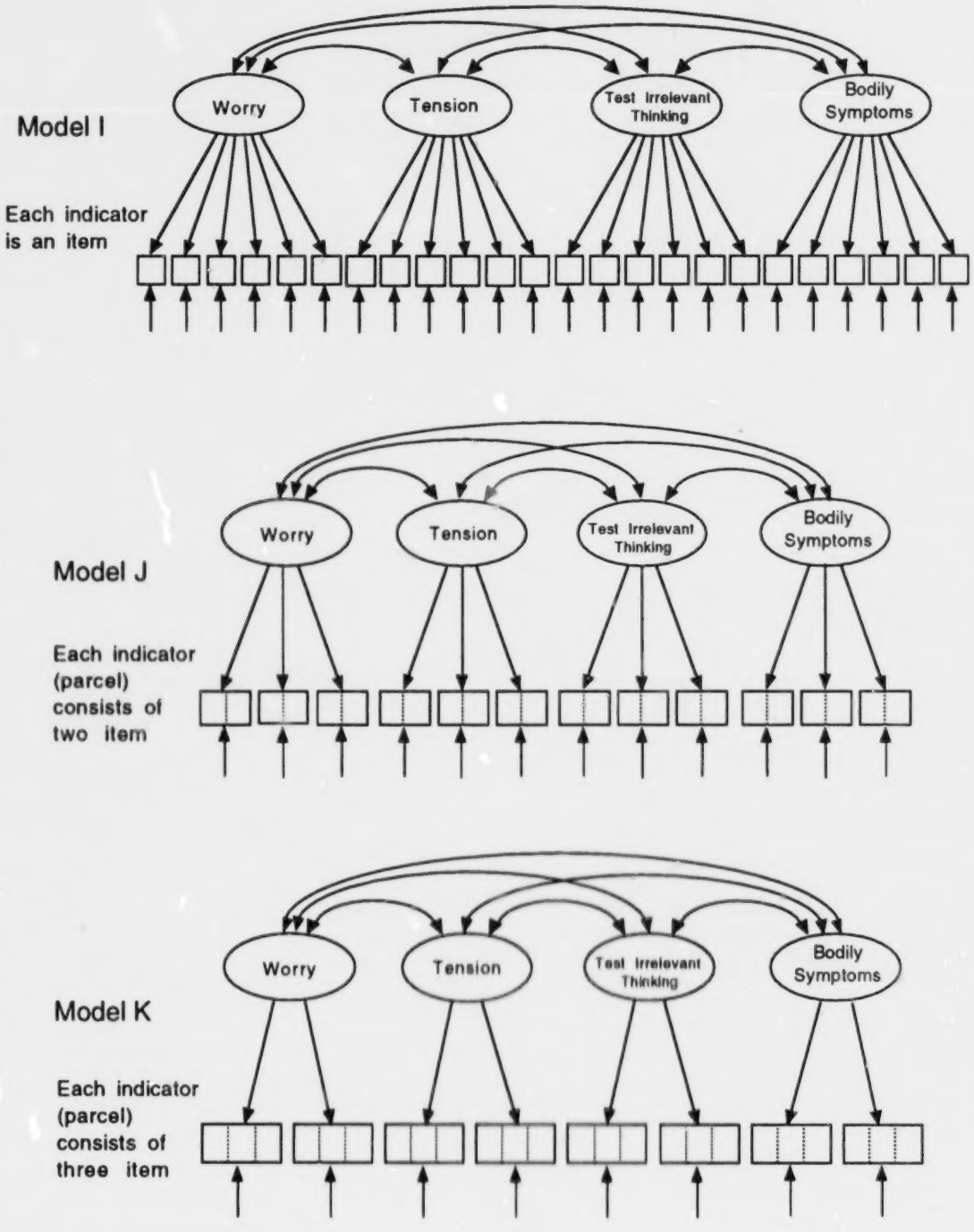

Figure 1. Example of three models with different numbers of parcels per factor and different numbers of items per parcel.

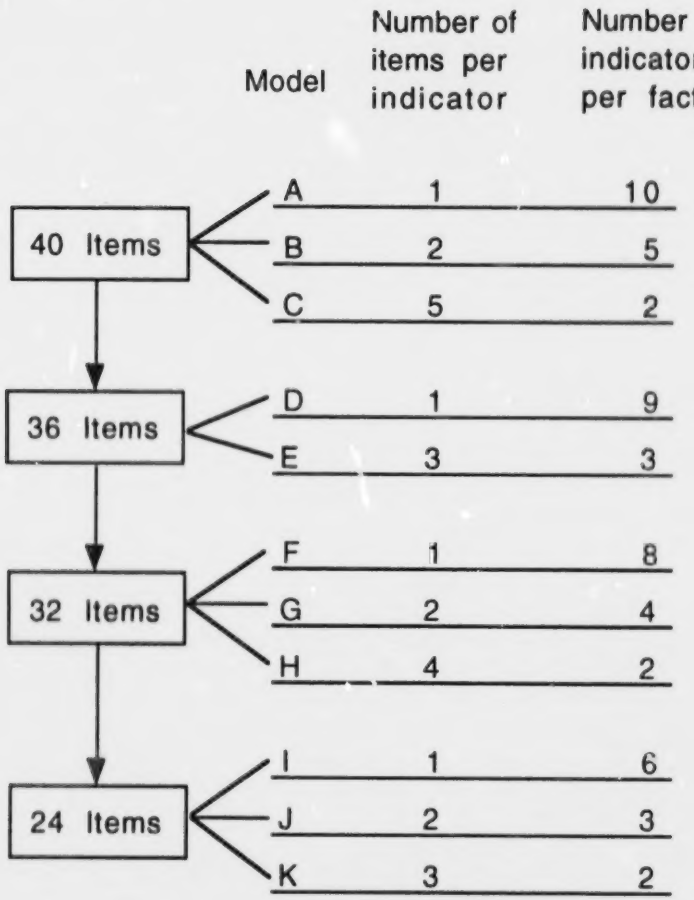

Figure 2. Description of the research design

Note. Each model consists of four factors

Criteria for Item Selection and Parceling

1. Item content
2. Inter-item correlation
3. Item reliability
4. Exploratory Factor Analysis (Factor loading)